D-optimal Designs and Equidistant Designs for Stationary Processes

Milan Stehlík

Department of Applied Statistics

Johannes Kepler University in Linz

Austria

Outlook

1. Correlated design and D-optimality

2. Equidistant designs for stationary processes

3. D-optimal designs for stationary processes

The isotropic stationary process

$$Y(x) = \theta + \varepsilon(x).$$

The design points $x_1, ..., x_N$ are taken from a compact design space $X = [a, b], \ b - a > 0$.

The mean parameter $E(Y(x)) = \theta$ is unknown.

The variance-covariance structure $C(d, r)$ depends on another unknown parameter r.

Fisher information matrices

In such model we have Fisher information matrices

$$M_\theta(n) = 1^T C^{-1}(r) \, 1$$

and (see Pázman (2004) and Xia et al. (2006))

$$M_r(n) = \frac{1}{2} tr \left\{ C^{-1}(r) \frac{\partial C(r)}{\partial r} C^{-1}(r) \frac{\partial C(r)}{\partial r^T} \right\}.$$

For both parameters of interest $M(n)(\theta, r) = \begin{pmatrix} M_\theta(n) & 0 \\ 0 & M_r(n) \end{pmatrix}.$

Regularity assumptions on covariance structure

a) $C(d,r) > 0$ for all r and $0 < d < +\infty$,

b) for all r is mapping $d \to C(d,r)$ continuous and strictly decreasing on $(0,+\infty)$

c) $\lim_{d \to +\infty} C(d,r) = 0$.

Example 1.

The power exponential correlation family $C(d, r) = \sigma^2 \exp(-rd^p)$, $0 < p \le 2$, $r > 0$. This family is by far the most popular family of correlation models in the computer experiments literature (see Santner et al. (2003)). The exponential $\exp(-rd)$ and Gaussian correlation functions $\exp(-rd^2)$ are special cases of the power exponential correlation family.

Example 2.

The Matérn class of covariance functions

$$\mathrm{cov}(d, \phi, v) = \frac{1}{2^{v-1}\Gamma(v)}\left(\frac{2\sqrt{v}d}{\phi}\right)^v K_v\left(\frac{2\sqrt{v}d}{\phi}\right)$$

(see e.g. Handcock and Wallis (1994)). Here ϕ and v are the parameters and K_v is the modified Bessel function of the third kind and order v.

The class is motivated by **(a)** the smoothness of the spectral density, **(b)** the wide range of behaviors covered, **(c)** and the interpretability of the parameters.

It includes the exponential correlation as a special case with $v = 0.5$ and the Gaussian correlation function as a limiting case with $v \to \infty$.

Which design is good?

We employ the D-optimality: to maximize the determinant of Fisher information matrix (FIM)

Classical interpretation: (Fedorov, Pázman, Pukelsheim) D-optimum design minimizes the volume of the confidence ellipsoid for θ.

We need justification of D-optimality under correlation, since

Correlation may lead to unexpected, counter-intuitive even paradoxical effects in the design (e.g. Müller and Stehlík, 2004 & 2007) as well as the analysis (e.g. Smit, 1961) stage of experiments.

Justification of D-optimality under correlation

a) The inverse of the FIM may well serve **as an approximation of the covariance matrix of maximal likelihood estimators** in special cases (Pázman (2004), Abt and Welch (1998), Zhu and Stein (2005), Zhang and Zimmerman (2005)).

b) Although some simulation and theoretical studies shows the limits of such an approximation of the covariance matrix of the ML estimates, it can still be used as a design criterion **if the relationship between these two are monotone**, since for the purpose of optimal designing **the only correct ordering is important**. For instance, Zhu and Stein (2005) observes a monotone relationship between them.

$M_\theta(n)$ **structure**

Example 1: *Exponential covariance structure* $\exp(-rd)$

For the sake of simplicity and without the loss of generality we fix $r = 1, X = [-1, 1]$. We have

$$M_\theta(2) = \frac{2e^d}{1 + e^d}$$

The D-optimal design is the maximal distant one.

If we consider three-point-design with distances $d_i = x_{i+1} - x_i, i = 1, 2$ then information

$$M_\theta(3) = 1 + \frac{2 + 2e^{-d_1 - 2d_2} - 2e^{-d_1} + 2e^{-2d_1 - d_2} - 2e^{-2(d_1 + d_2)} - 2e^{-d_2}}{e^{-2(d_1 + d_2)} - e^{-2d_1} - e^{-2d_2} + 1}.$$

In Stehlík (2004) is proved that $\{-1, 0, 1\}$ is D-optimal design for θ.

The complexity of $M_\theta(n)$ increases significantly with n (see Stehlík (2007)).

The equidistant designs

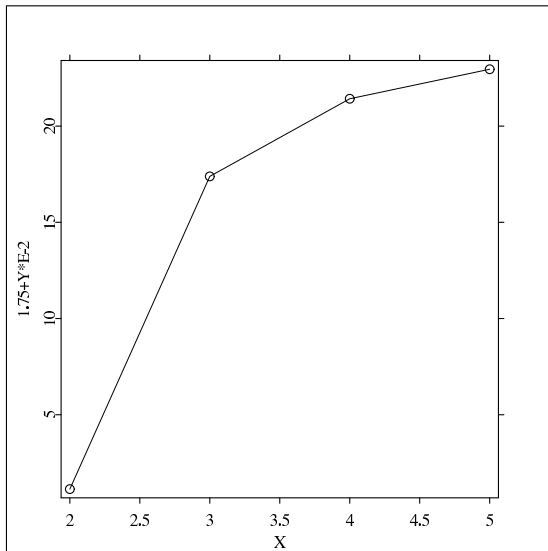The structure of $M_\theta(n)$ for equidistant designs is much more simple, i.e.

$$M_\theta(2) = \frac{2e^d}{1 + e^d}, \ M_\theta(3) = \frac{-1 + 3e^d}{1 + e^d}, \ M_\theta(4) = \frac{-2 + 4e^d}{1 + e^d}.$$

Kiseľák and Stehlík (2007):

$$M_\theta(k) = \frac{2 - k + ke^d}{1 + e^d}.$$

Note that $\lim_{d \to +\infty} M_\theta(k)/M_\theta(k - 1) = \frac{k}{k-1}, \ k \geq 3.$

$M_\theta(n)$ **is increasing with number of the design points**

Theorem 4 in Kiseľák and Stehlík (2007): *The equidistant design for parameter θ is D-optimal.*

This theorem is some extension of the **Theorem 3.6** in Dette, Kunert and Pepelyshev (2006). Therein is proved, that for $r \to 0$ the exact n-point D-optimal design in the linear regression model with exponential covariance converges to the equally spaced design.

A lower bound for $M_\theta(n)$

$$LB(d) := n \inf_x \frac{x^T C^{-1}(d,r) x}{x^T x}.$$

Theorem 1 in Stehlík (2007).

Let $C(d,r)$ is a covariance structure satisfying a),b),c). Then

1) for any design $\{x, x + d_1, x + d_1 + d_2, ..., x + d_1 + ... + d_{n-1}\}$ given by distances $d_i, i = 1, ..., n-1$ and for any subset of distances $d_{i_j}, j = 1, ..., m$ the lower bound function $(d_{i_1}, ..., d_{i_m}) \to LB(d)$ is increasing in the d's. In particular, for any equidistant design ($\forall i : d_i = d$) the function $d \to LB(d)$ is increasing in d.

2) $\lim_{\forall i: d_i \to +\infty} M_\theta(n)/M_\theta(n-1) = \frac{n}{n-1}$.

The proof is based on Frobenius Theorem and smoothness of matrix inverse.

Illustrative example

Let us consider a *power exponential covariance* family with zero nugget. For the sake of simplicity let us consider the equidistant designs. We have

$$M_\theta(2) = \frac{2e^{rd^p}}{1 + e^{rd^p}}, \; M_\theta(3) = \frac{e^{d^p r}\left(e^{d^p r} - 4e^{2^p d^p r} + 3e^{(1+2^p)d^p r}\right)}{e^{2d^p r} - 2e^{2^p d^p r} + e^{(2+2^p)d^p r}}.$$

Then $\lim_{d \to +\infty} M_\theta(k)/M_\theta(k-1) = \frac{3}{2}$.

Estimation of the covariance parameter r.

Problem: *One of the fundamental assumptions, the knowledge of the covariance function, is in most cases almost unrealistic. "It seems to be artificial, that the first moment $E(Y(x))$ is assumed to be unknown whereas the more complicated second one is assumed to be known..." Näther (1985)*

M_r is much more complex than M_θ

Example: $M_r(2) = \frac{d^2 \exp(-2rd)(1+\exp(-2rd))}{(1-\exp(-2rd))^2}$.

Two point optimal design for parameter r is collapsing ($d_{D-optimal} = 0$)! This also holds for pair (θ, r)!

Nugget effect

The distance of the two point D-optimal design for covariance parameter r of exponential covariance can be tuned by the **nugget**

$$\tau^2 = \lim_{d \to 0} \frac{1}{2} Var(Y(x + d) - Y(x))$$

In Stehlík, Rodríguez-Díaz, Müller and López-Fidalgo (2007) is proved that **the distance of D-optimal design is an increasing function of the nugget** τ^2.

The complexity of $M_\theta(n)$ increases very significantly with n (see Stehlík (2007)).

For n-point equidistant design the relation $(n-1)M_r(2) = M_r(n)$ is proved in Kiseľák and Stehlík (2007).

However, if the nugget $\tau^2 > 0$ then $(n-1)M_r(2) \neq M_r(n)$, but equality can be reached in the limit $\tau^2 \to 0$, e.g.

$$\lim_{\alpha \to 1} M_{r,1-\alpha}(3)/M_{r,1-\alpha}(2) = 2 = M_r(3)/M_r(2).$$

Comparisons of equally spaced designs

Hoel (1958) provided asymptotical comparisons made for equally spaced sets of points. The sets of points that he selected for consideration were the following:

(a) n equally spaced points in the interval $(0, l)$

(b) $2n$ equally spaced points in the interval $(0, l)$

(c) $2n$ equally spaced points in the interval $(0, 2l)$

(d) two sets of observations of type (a)

We consider ratios:

$$R_1 = \left[\frac{M(n,d)}{M(2n,d/2)}\right]^{-1}$$

$$R_2 = \left[\frac{M(n,d)}{M(2n,d)}\right]^{-1}$$

$$R_3 = \left[\frac{M(n,d)}{M(m\ runs)}\right]^{-1}.$$

These ratios are used in the three cases,

(A) when the only trend parameter θ is estimated, **(B)** when the only correlation parameter r is estimated, **(C)** when the both parameters (θ, r) are estimated.

In Kiseľák and Stehlík (2007) we have observed the following results:

(a) for all possible combinations of parameters of interest, i.e. $\{\theta\}, \{r\}$ and $\{\theta, r\}$, the interval over which observations are to be made **should be extended as far as possible** *(increasing domain asymptotics)*

(b) However doubling the number of observation points in a given interval *(infill domain asymptotics)*, when the only parameter θ is of interest and there are already a large number of such points, gives practically no additional estimation information. When $\{r\}$ or $\{\theta, r\}$ are the sets of interest, doubling gives the double information.

References

Abt M. and Welch W.J. (1998). Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes, The Canadian Journal of Statistics, Vol. 26, No. 1, 127-137.

Dette, H., Kunert, J. and Pepelyshev, A. (2006). Exact optimal designs for weighted least squares analysis with correlated errors, accepted to Statistica Sinica.

Handcock M.S. and Wallis J.R. (1994). An Approach to Statistical Spatial-Temporal Modeling of Meteorological Fields, JASA, 89, 368-378.

Hoel P.G. (1958). Efficiency Problems in Polynomial Estimation. Annals of Math. Statistics, 29, 1134-1145.

Kiseľák J. and Stehlík M. (2007). Equidistant D-optimal designs for parameters of Ornstein-Uhlenbeck process, IFAS research report Nr. 22.

Müller W.G. and Stehlík M. (2004). An example of D-optimal designs in the case of correlated errors, Proceedings in Computational Statistics, Edited by J. Antoch, Physica-Verlag, 1543-1550.

Müller W.G. and Stehlík M. (2007). Issues in the Optimal Design of Computer Simulation Experiments, IFAS research report.

Näther W. (1985). Effective Observation of Random Fields, Teubner- Texte zur Mathematik, Band 72, BSB B. G. Teubner Verlagsgesellschaft, Leipzig.

Pázman A. (2004). Correlated optimum design with parametrized covariance function: Justification of the use of the Fisher information matrix and of the method of virtual noise, *Report #5 Research Report Series*, Department of Statistics and Mathematics, WU Wien.

Santner T.J., Williams B.J. and Notz W.I. (2003). The Design and Analysis of Computer Experiments, Springer, New York.

Smit J.C. (1961). Estimation of the mean of a stationary stochastic process by equidistant observations. Trabojos de estadistica 12, 34-45.

Stehlík M. (2004). Some Properties of D-optimal Designs for Random Fields with Different Variograms, *Report #4 Research Report Series*, Department of Statistics and Mathematics, WU Wien.

Stehlík M. (2007). D-optimal designs and equidistant designs for stationary processes, mODa 8 Proceedings, Eds. López-Fidalgo, J., Rodríguez-Díaz J.M. and Torsney, B. (Eds.), 2007.

Stehlík M., Rodríguez-Díaz J.M., Müller W.G. and López-Fidalgo J. (2007). Optimal allocation of bioassays in the case of parametrized covariance functions: an application in lung's retention of radioactive particles, *TEST* (in press).

Xia, G., Miranda, M.L. and Gelfand, A.E. (2006). Approximately optimal spatial design approaches for environmental health data, Environmetrics, 17, 363-385.

Zhang H. and Zimmerman D. (2005). Towards reconciling two asymptotic framework in spatial statistics, Biometrika, 92, 4, pp. 921-936.

Zhu, Z. and Stein, M.L. (2005), Spatial sampling design for parameter estimation of the covariance function, Journal of Statistical Planning and Inference, vol. 134, no 2, 583-603.

Thank you for your attention!